This is a pre-copyedited version of a contribution published in V. Müller (ed.), Philosophy and Theory of Artificial Intelligence 2017, published by Springer Nature Switzerland AG.

# Will machine learning yield machine intelligence?

Carlos Zednik carlos.zednik@ovgu.de
Otto-von-Guericke Universität Magdeburg, Germany

**Abstract** One way of going beyond the Turing Test is to consider whether Machine Learning (ML) methods will eventually yield algorithms that resemble the ones that are implemented in human brains. Unfortunately, the well-known *Black Box Problem* makes it difficult to know which algorithms these methods actually yield. Recent attempts to solve this problem within *Explainable AI* may therefore deliver an *a posteriori* answer to the philosophical question of machine intelligence. That said, *a priori* considerations also suggest that eventually, ML-programmed computers will not only behave like humans, but will also acquire algorithms that are similar to the ones that are implemented in human brains.

## 1. The Algorithmic Similarity Criterion for Machine Intelligence

The recent successes of Machine Learning (ML) spark renewed interest in the central philosophical question of Artificial Intelligence: *Can machines think*? Rather than program computers with hand-coded algorithms for solving complex problems, these methods allow computers to "program themselves", by learning such algorithms through e.g. the identification of regularities in large bodies of data, or through extensive trial and error in a real or simulated environment. In this way, ML methods have already produced self-driving cars and autonomous helicopters, computers that play chess and Go, and sophisticated face-, handwriting-, and speech-recognition systems, among many others.

These technological advances are undeniably impressive feats of engineering; but do they also herald an age of thinking machines? Even if ML-programmed computers are eventually shown to match or exceed human behavior in a variety of domains, several commentators have argued that more is required to demonstrate a genuine capacity for intelligent thought. In order to preclude "pretense"—e.g. computers that act like humans but whose behavior is governed by a lookup table (Block, 1981)—it may be necessary to "look under the hood" by considering which algorithms drive a machine's behavior. That is, a suitable criterion for machine intelligence may be *Algorithmic Similarity* (AS): A computer is intelligent if it operates according to rules and representations similar to the ones that are implemented in human brains. Notably, although AS may be a sufficient criterion for machine intelligence, it is hardly necessary; although a computer whose programming resembles our own would be considered intelligent, a genuinely intelligent computer might also be programmed quite differently.

## 2. The Black Box Problem and Explainable AI

Unfortunately, it is difficult to know which algorithms are acquired in the process of Machine Learning, and thus, whether any particular ML-programmed computer actually satisfies AS. Indeed, the high-dimensional complexity of deep neural networks and sophisticated reinforcement learning policies are such that their inner workings often remain opaque to human users—the so-called *Black Box Problem* in AI. The Black Box Problem is already known to have considerable practical ramifications (e.g. Ribeiro, Singh, & Guestrin, 2016); it now appears to have philosophical significance as well. Can the Black Box Problem be solved?

An *a posteriori* solution to the Black Box Problem may eventually be provided by the nascent *Explainable AI* research program (Doran, Schulz, & Besold, 2017). One prominent branch of Explainable AI aims to reduce the opacity of ML-programmed computers through mathematical tools, experimental techniques, and visualization methods that characterize the algorithms that drive deep neural networks and other ML-programmed computers (e.g. Ritter, Barrett, Santoro, & Botvinick, 2017). Another important branch of Explainable AI aims to modify the Machine Learning process itself so that computers do not only learn to solve complex problems, but so that they also learn to provide comprehensible "explanations" of their actions (e.g. Ribeiro et al., 2016).

Of course, Explainable AI's ability to solve the Black Box Problem—and thus, its ability to inform the philosophical questions at hand—remain uncertain. For one, Explainable AI may fail to reveal the algorithms that are acquired through Machine Learning. For another, Explainable AI may solve the Black Box Problem, but reveal that ML-programmed computers are algorithmically quite *dis*-similar to intelligent human beings. In either one of these cases, although Explainable AI would yield no positive evidence for machine intelligence, it would yield no negative evidence, either; we would simply be unable to tell whether ML-programmed computers are genuinely intelligent by "looking under the hood". Alternatively, Explainable AI may also succeed, and reveal that (some) ML-programmed computers do in fact satisfy the Algorithmic Similarity criterion. In this case, the research program would have provided *a posteriori* reasons to believe in the existence of genuinely intelligent machines.

## 3. "Nurture" and Situatedness

But just how likely is this third outcome? *A priori*, it may appear unreasonable to expect Machine Learning to yield algorithms that resemble the ones that are implemented in human brains. Given that many learning algorithms such as backpropagation are biologically implausible (Bengio, Lee, Bornschein, Mesnard, & Lin, 2015)—and thus, that ML-programmed computers' "nature" is quite unlike the nature of intelligent human beings—it may seem unreasonable to expect the algorithms that are acquired through Machine Learning to resemble the ones that are implemented in human brains. Nevertheless, there are at least two reasons to think otherwise.

First, although their "nature" may be quite different, their "nurture" is increasingly similar. Today, many ML-programmed computers have access to the

very same information that is also given to human children. This includes text produced by humans *for* humans (e.g. newspaper articles, books, and Twitter feeds), as well as highly naturalistic images, sounds, and videos (e.g. the ones that appear on Instagram and YouTube). Insofar as the algorithms acquired through Machine Learning are more determined by a computer's real-world "nurture" than by its artificial nature, there is reason to believe that ML methods will not only yield computers that behave like humans, but that these computers will also be governed by algorithms that resemble the ones that acquired in regular human education.

Second, ML-programmed computers are increasingly situated in the same real-world environment that is also inhabited by humans. Self-driving cars must learn to navigate not only in simulation, but also on the same roads that are used by human drivers. As a consequence, the former will likely learn to exploit the same structures, artifacts, and tools that are already being exploited by the latter. For example, a self-driving car might learn to exploit the presence of road signs to navigate in much the same way that humans do, and in this sense, will acquire methods of interacting with the environment that closely resemble our own. Indeed, insofar as environmental structures, artifacts and tools might even be said to co-constitute "extended" human cognitive systems (Clark & Chalmers, 1998), they can also be said to co-constitute "extended" artificial cognizers. In this sense, ML-programmed computers may not only exhibit Algorithmic Similarity, but may even exhibit a degree of Algorithmic *Equivalence*.

In summary, although it's prospects are uncertain, Explainable AI may eventually provide an *a posteriori* answer to the question of whether ML-programmed computers satisfy AS. In addition, there are *a priori* reasons to believe that Machine Learning will not only yield computers that behave like humans in a variety of domains, but that these computers will in fact be driven by algorithms that resemble the ones that are implemented in human brains. Insofar as this is a sufficient criterion for machine intelligence, Machine Learning may in fact bring about an age of thinking machines.

## References


Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv*, 1502.04156.

Block, N. (1981). Psychologism and Behaviorism. *The Philosophical Review*, *90*(1), 5–43.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 7–19.

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *arXiv*, 1710.00794.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv*, 1602.04938v3.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. In *Proceedings of the 34th International Conference on Machine Learning*.